Question Answering with Attentions Ensemble

Zihan Lin[‡], Teng Zhang[†], Jason Zhu[†]
[†] Department of Management Science and Engineering, Stanford University
[‡] Institute for Computational and Mathematical Engineering, Stanford University

Introduction

In this project, our group develops a Deep Learning system for the Stanford Question Answering Dataset (SQuAD), which is a large-scale dataset for reading comprehension and question answering created through crowd-sourcing. Given a context paragraph and a question about that paragraph, our model aims to correctly predict the answer of the question, which is assumed to be a continuous sub-span of the context paragraph. Under this setting, the goal of the model is essentially to correctly predict the starting and ending index of the answer in the context.

Model Structure

BiDAF

Two features:

- The attention vector at each time step, along with the embeddings from previous layers, are allowed to flow through to the subsequent modeling layer
- The attention is computed in two directions: from context to query as well as from query to context

Coattention

We utilize the two of the contributions in the original paper. One is introducing the two-way attention structure into our model. The other is adding the sentinel vectors.

R-Net

The third attentive model we replicated is the R-Net model proposed by Microsoft Research Asia. We followed the structure of the original model, but with some small tweaks. We first feed word and character level embeddings into bi-directional RNN to produce new representations of the words in both question and context. A Gated Attention layer is then used to match question and context. After is a self-attention layer to learn passage context. Lastly, a pointer network is used to get the starting and ending index of the answer.

Ensemble

We first take the output probabilities of the starting and ending index of the answers from each model's output layer, and then take the average of the probability over three models. lastly, argmax of the starting and ending index is taken to generate the final answer.

Experiment

Implementation Details

BiDAF

- Character embedding trained on CNN. Combined with Glove embedding, they are transferred to highway network.
- The dimension of embedding matrix is lower than the original paper.

Coattention

Based on the relatively fast training of the model, we perform parameter tuning sequentially of the following parameters: *learning_rate, hidden_size, embedding_size, drop_out, context_len*.

R-Net

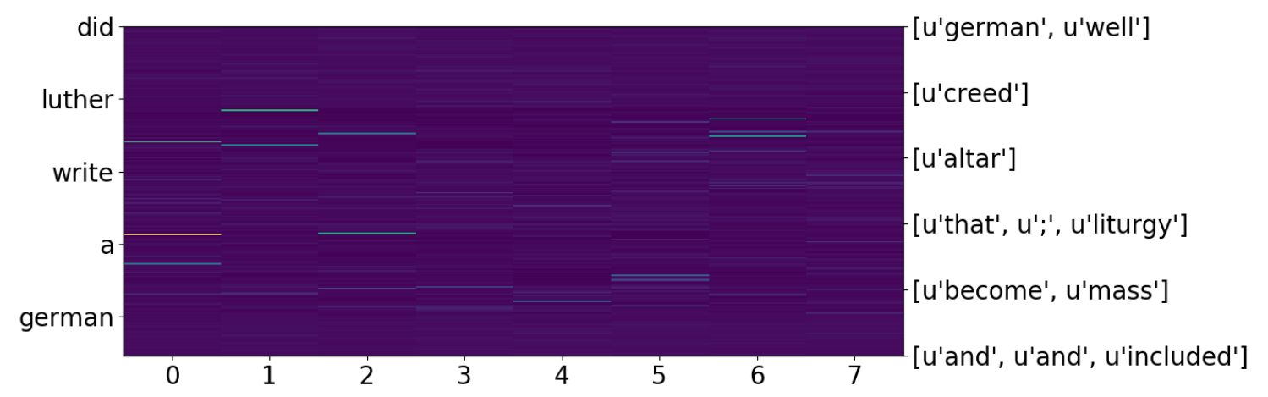
The word-level and character-level embeddings we used are both pretrained GloVe vectors. To speed up calculation, scaled multiplicative attention proposed in *Attention is All You Need* is used. All bi-directional RNN use GRU cells. Most parameters are chosen to match the parameters reported in the original R-Net paper.

Main Results

Model Performance

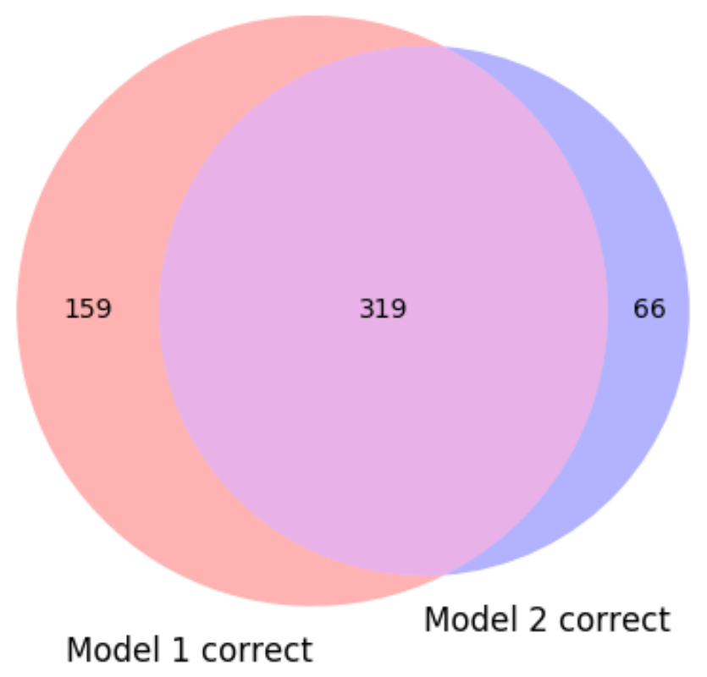
Method	F_1	EM
R-Net	0.66	0.59
Coattention	0.57	0.48
BiDAF	0.65	0.59

Visualization of the attention matrix for an example

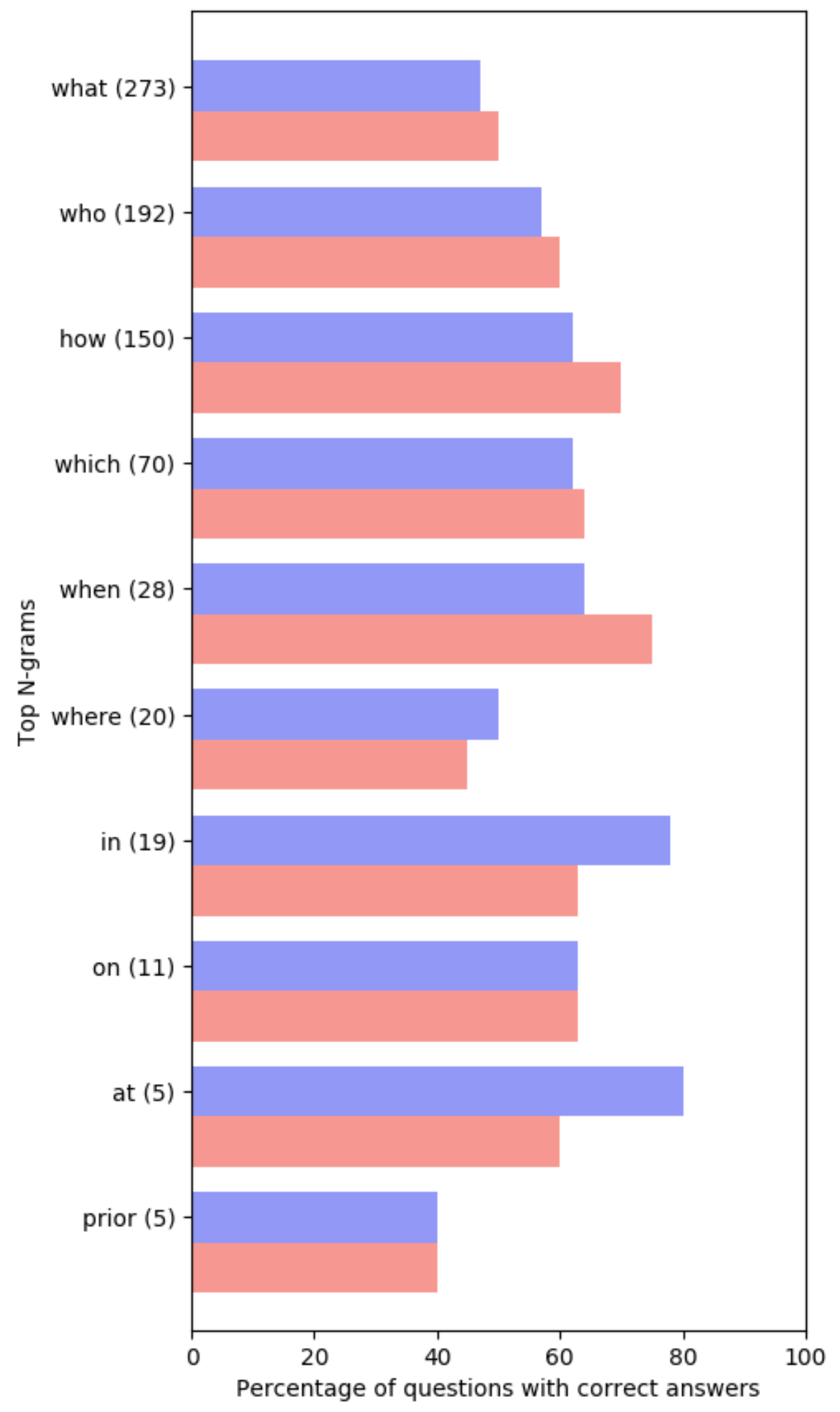


We visualize the attention matrices for some question-context tuples in the dev data in figure. We find that entities in the question typically attend to the same entities in the context, thus providing a feature for the model to localize possible answers.

Comparison between different models



Venn diagram of the questions answered correctly by BiDAF (Model 2) and R-Net (Model 1)



Correctly answered questions broken down by the 10 most frequent first words in the question, BiDAF (Blue bar) and R-Net (Red Bar)

Error analysis of some examples

Error type	Example
Imprecise answer boundaries	**Context:** ...during that year , tesla worked in pittsburgh , helping to create an alternating current system to power the city 's streetcars ... **Question**: what did tesla work on in 1888 ? **True Answer**: system to power the city 's streetcars **Predicted Answer**: helping to create an alternating current system to power the city 's streetcars
Syntactic complications and ambiguities	**Context:** ...if the head of government of a country were to refuse to enforce a decision of that country 's highest court , it would not be civil disobedience... **Question**: what does not constitute as civil disobedience ? **True Answer**: refuse to enforce a decision **Predicted Answer**: refuse to enforce a decision
External knowledge	**Context:** ...atp synthase uses the energy from the flowing hydrogen ions to phosphorylate adenosine diphosphate into adenosine triphosphate , or atp . because chloroplast atp synthase projects out into the stroma , the atp is synthesized there... **Question**: what does atp synthase change into atp ? **True Answer**: phosphorylate adenosine diphosphate **Predicted Answer**: stroma

Conclusion

All of our three models obtain superior performance to the original baseline model, and does reasonable performance on most of the errors they make (see error analysis). By comparing three models, we see the advantages and disadvantages of different models, and gain intuitive understanding of why these models work so well in question answering tasks.

Future Works

Ensemble is a quick and naive way to combine multiple models. If we have more time, we can try to combine layers and great ideas from different papers into a single model, which might have better performance and interpretability.